УДК 004.04.043; 004.912; 004.62

К.К. Духновська
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України
Україна, 03680, м. Київ, пр. Глушкова, 40

ФОРМУВАННЯ ПОШУКОВОГО ДИНАМІЧНОГО ВЕКТОРНОГО ПРОСТОРУ

K.K. Duchnovska
International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine
40 Glushkova ave., Kyiv, Ukraine, 03680

FORMATION OF THE RESEARCH DYNAMIC VECTOR SPACE

К.К. Духновская
Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины
Украина, 03680, г.Киев, пр. Глушкова, 40

ФОРМИРОВАНИЕ ПОИСКОВОГО ДИНАМИЧЕСКОГО ВЕКТОРНОГО ПРОСТРАНСТВА

У статті обґрунтовується подання текстового документа у векторному вигляді для подальшого застосування алгебраїчного апарату в алгоритмах пошуку інформації. Текстовий документ представляється *TF-IDF*моделлю, в яку введено динамічну складову.
Ключові слова: текстовий пошук, *TF-IDF*модель, пошуковий векторний простір.

The article substantiates the idea of text document vector for further use in the apparatus of algebraic algorithms for searching information. Text Document appears as TF-IDF model in which dynamic component is introduced
Keywords: text search, TF-IDF model, the search vector space.

В статье обосновывается представление текстового документа в векторном виде для дальнейшего применения алгебраического аппарата в алгоритмах поиска информации. Текстовый документ представляется *TF-IDF* моделью, в которую введено динамическую составляющую.
Ключевые слова: текстовый поиск, *TF-IDF* модель, поисковое векторное пространство.

Вступ

Зростання матеріальних і духовних цінностей людства, темпів розвитку науки і техніки знаходить своє відображення у великій кількості не структурованих документів, що заповнюють простір сучасних інформаційних сховищ. Основна частина інформації (близько 80%) представлена в текстовому вигляді. Тому проблематика текстового пошуку є особливо актуальною.

Для побудови алгоритмів текстового пошуку активно застосовується математичний апарат. Але представлення текстового документа як вектора не сприймається, а іноді заперечується. Тому дуже важливо розглянути дане представлення з позиції векторної аксіоматики.

Одночасно, у всіх відомих моделях текстового документа використовуються статичні елементи, що не відповідає дійсності. Тому як актуальність документів та інформації взагалі змінюється з часом. Впровадження динаміки в елементи моделі текстового документа є необхідним кроком для покращення якості пошуку і оптимізації всього пошукового процесу.

Для викладення матеріалу статті наведемо деякі відомі визначення.

Визначення 1. Під текстом розуміють кінцеву множину слів, які утворюють інформативне повідомлення і об'єднані лексичним, граматичним, змістовним і частотним співвідношенням.

Визначення 2**.** Інформаційним пошуком називають процес, в результаті якого відбувається виявлення потрібної інформації в деякій множині текстових документів, фактів і т.д..

Інформаційними ресурсами (ІР) будемо називати документи подані в електронному вигляді.

Накопичення інформації. На рис. 1 схематично представлено процес накопичення інформаційних ресурсів в електронному сховищі. Під електронним сховищем розуміється довільне файлове сховище текстових ІР.

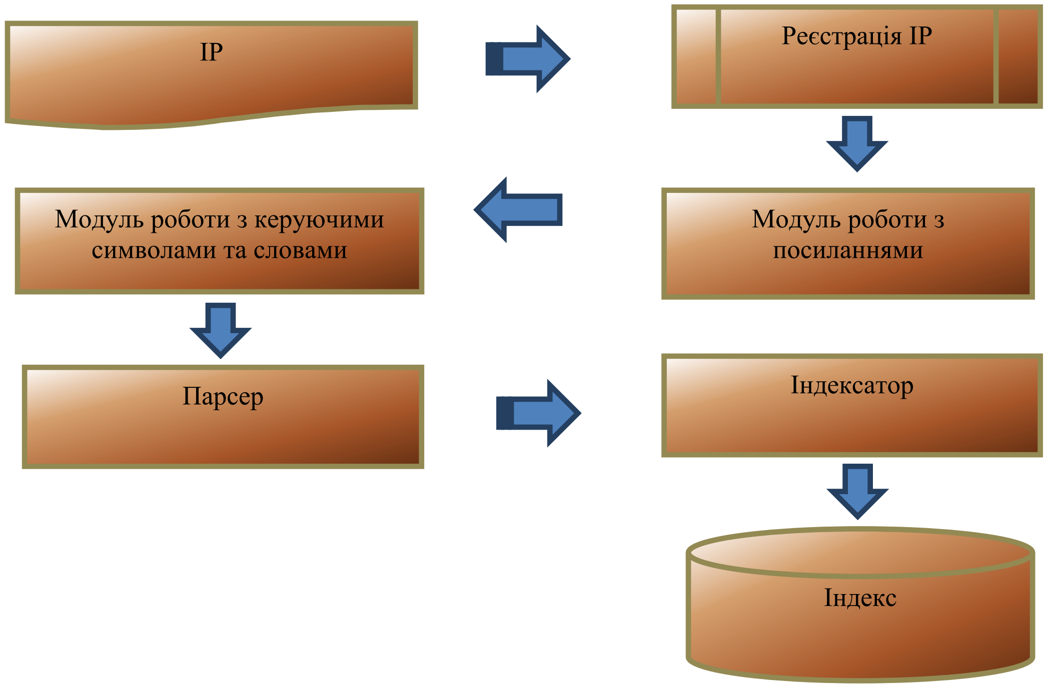


Рис.1. Первісна обробка ІР

У першу чергу ІР реєструється в базі електронного сховища. Після реєстрації ІР надходить на виділення посилань. Посилання з поточного ІР потрапляють в чергу для завантаження з цієї адреси нового ІР.

Далі ІР надходить в модуль, в якому видаляються з ІР керуючі символи, команди і т.п. На виході отримуємо текст ІР без усього зайвого, який передається до парсеру – спеціального модуля, функцією якого є синтаксичний аналіз тексту та виокремлення слів з тексту ІР. Даний модуль розраховує всі метрики, які необхідні для класифікації ІР та його пошуку. Потім ІР, або множина відокремлених термінів з нього подається в індексатор, який класифікує ІР, тобто

знаходить належне йому місце в категоріях електронного сховища, і записує у відповідному форматі.

Кожен текстовий ІР – це сукупність термінів, яка несе деяку інформацію. Термін – це синтаксично самостійний комплекс морфем, що утворюють жорстко пов'язану структуру. Термін відрізняється від поєднання слів тим, що деякі його елементи не можуть вживатися в синтаксично ізольованій позиції. Крім того, елементи всередині терміну пов'язані один з одним набагато більш жорсткими і міцними зв'язками, ніж елементи речення (тобто поєднання слів). Щоб врахувати всі словоформи окремого терміну застосовуються алгоритми лемматізації і стеммінгу.

Лемматізація – це приведення різних форм термінів у відповідність з граматичними формами певної мови.

Стеммінгом називають наближений евристичний процес, на вході якого від слів відкидаються закінчення в розрахунку на те, що в більшості випадків це себе виправдає, тобто мається на увазі видалення похідних афіксів. Із застосуванням механізмів стеммінга з'являється можливість робити пошук ІРз урахуванням морфології слова. Це означає, що при введенні користувачем запиту, враховуються всі словоформи даного терміну.

На сьогодні існує багато різноманітних алгоритмів, які впроваджують стеммінг. Серед них виділяють стреммер Портера, алгоритми *KSTEM* і*n*-грам. Алгоритм Портера не використовує баз основ слів, а лише, застосовуючи послідовно ряд правил, відсікає закінчення і суфікси, ґрунтуючись на особливостях мови, у зв'язку з чим працює швидко, але не завжди безпомилково. Перевагою алгоритму *KSTEM* є те, що він не залежить від частини мови терміну, а спирається на алгоритм заміни суфікса. Алгоритм *n* – грам ґрунтується на принципі: «Якщо слово *A* збігається зі словом *B* з урахуванням декількох помилок, то з великою часткою ймовірності в них буде хоча б один спільний підрядок довжиною *N*». Ці підрядки довжиною *N* і називаються *n*-грамами. Під час індексації слово розбивається на такі *n*-грами, а потім це слово потрапляє в списки для кожної з цих *N*-грам. Під час пошуку запит також розбивається на *n*-грами, і для кожної з них проводиться послідовний перебір списку термінів, що містять даний підрядок [1].

Моделі ІР

Під моделлю ІР розуміють сукупність будь-яких характеристик ресурсу, які враховуються системою при його обробці. Характеристики ІР поділяють на два типи: пов'язані з текстом ІР і непов'язані з текстом – атрибути ІР. До характеристик, пов'язаних з текстом, відносять присутність термінів, їх розташування в тексті відносно один одного, форматування документа, структура ІР. Характеристики, не пов'язані з текстом, в системах *Web*-пошуку називаються «мета-атрибутами». Такі атрибути беруться з інших джерел. Для цього виду пошуку як атрибути використовують *URL*-адресу ІР в мережі *Internet*, інформацію про час створення або зміни ресурсу.

У моделях ІР, характеристики яких пов'язані з текстом, у простому випадку розглядається тільки факт наявності або відсутності слів у документі. Таку модель ІР називають бінарною. Більш удосконаленим варіантом такої моделі є підхід, де для кожного терміну вказується не тільки його наявність, але і деяка "вага".

Найбільш поширеними методами зважування термінів в ІР, пов'язані з отриманням наступних характеристик:

1) кількістю появ термінів у даному ІР. Дана характеристика досить проста й очевидна. Якщо термін частіше міститься в тексті ІР, то, швидше за все, цей ІР більш пов'язаний за змістом з цим терміном. Недоліком цього методу оцінки "ваги" є те, що якщо колекція містить ІР різної довжини, то більшу вагу будуть отримувати більш довгі ресурси, так як в них більше термінів;

2) частотою появи термінів в ІР (*TF*). Дана характеристика обчислюється як відношення числа входження терміну до загальної кількості термінів ІР. Недоліком є те, що в даному випадку, навпаки, недооцінюються довгі документи, так як в них більше термінів і їх середня частота в тексті ІР нижча. Для вирішення цієї проблеми застосовується доповнена нормалізована частота, яка обчислюється як 0.5+0.5(*TF*/*ATF*), де *ATF*-середня частота терміна в електронному сховищі;

3) логарифмом частоти входження терміну. У даному випадку вага терміну, що входить в текст ІР визначається як *1+log(TF),* де *TF* - частота терміна. Для компенсації ефекту різної довжини ресурсів використовують аналогічну нормалізацію частоти. У цьому випадку формула виглядає як *(1+log(TF)/(1+log (MTF)),* де *MTF*-максимальна частота терміну в електронному сховищі ІР.

Експериментально доведено, що урахування ваги документа на підставі статистичних характеристик покращує якість пошуку. Практично всі сучасні пошукові системи використовують одну з описаних характеристик, в основному варіанти використання частоти терміну в тексті ІР (*TF*).

Пошуковий векторний простір

Нехай маємо словник – упорядкований набір термінів, потужність якого *M*. Потужність словника – це кількість термінів, які в ньому містяться.

Після первинного опрацювання ІР (рис. 2) можна представити:

$$Di = \langle w_1, w_2, \ldots, w_{Mi} \rangle, \tag{1}$$

де w_k– частота терміна *k*-ого терміну(i=$\overline{1,M}$);

W – словник.

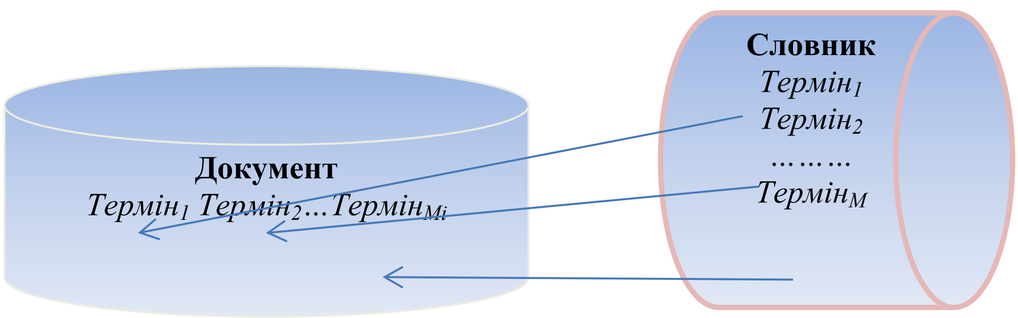


Рис. 2. Витяг термінів з документа

Нехай частота терміна розраховується за формулою *TF-IDF*:

$$TF = \frac{m_{ki}}{M_i}, \tag{2}$$

де: m_{ki} – кількість входжень k-ого терміну в i-ий ІР;
M_i – загальна кількість термінів в вi-омуІР;

$$IDF = ln(\frac{N}{N_k}) \quad (3)$$

де:N – загальна кількість ІР в електронному сховищі;
N_k – кількість ІР, в яких зустрічається k-ий термін.

Тоді:

$$w_k = TF * IDF \quad (4)$$

Доведемо, що представлення (1) є вектором.

Згідно з визначенням, вектором називається сукупність дійсних чисел, розташованих у певному порядку[2]. Представлення (1) відповідає визначенню, тому як кожна координата t_k займає місце відповідне розташуванню у словнику W. Для представлення (1) зберігаються всі векторні аксіоми.

1.Сума двох векторів в даному випадку – це злиття двох ІР:
$D^1 + D^2 = D^2 + D^1 =< w_1^1 + w_1^2, w_2^1 + w_2^2, ..., w_M^1 + w_M^2 >$ – комутативність складання.

2.$(D^1 + D^2) + D^3 = D^1 + (D^2 + D^3)$
$=< w_1^1 + w_1^2 + w_1^3, w_2^1 + w_2^2 + w_2^3, ..., w_M^1 + w_M^2 + w_M^3 >$ – асоціативність складання.

3. Нехай λ – скаляр. Добуток ІР на скаляр – це тиражування цього ресурсу скаляр разів.
$\lambda(D^1 + D^2) = \lambda D^1 + \lambda D^2 = \; =< \lambda w_1^1 + \lambda w_1^2, \lambda w_2^1 + \lambda w_2^2, ..., \lambda w_M^1 + \lambda w_M^2 >$ – дистрибутивність добутку відносно суми.

4. Нехай μ - скаляр, тоді:
$(\lambda + \mu)D = \lambda D + \mu D = \; < (\lambda + \mu)w_1^1, (\lambda + \mu)w_2^1, ..., (\lambda + \mu)w_M^1 >$

5.$\lambda(\mu D) = (\lambda\mu)D = \; < \lambda\mu w_1^1, \lambda\mu w_2^1, ..., \lambda\mu w_M^1 >$ - асоціативність добутку.

6.$\vec{0}$ – нульовий вектор: <0, 0, …, 0> – порожній ІР. Тоді:
0*D= $< 0 * w_1^1, 0 * w_2^1, ..., 0 * w_M^1 >$ =<0,0,…,0>=$\vec{0}$ – добуток будь-якого вектора на 0 є нульовий вектор – порожній ІР.

7. 1*D= $< 1 * w_1^1, 1 * w_2^1, ..., 1 * w_M^1 > = < w_1^1, w_2^1, ..., w_M^1 >$ =D– добуток будь-якого вектора на 1 дорівнює тому ж самому вектору.

Віднімання двох векторів визначається через добуток на -1 і формулою для суми: $D^1 - D^2 = D^1 + (-1)D^2$

Тобто: $D^1 - D^2 = \; < w_1^1 - w_1^2, w_2^1 - w_2^2, ..., w_M^1 - w_M^2 >$ і тоді виходить, що віднімання є дія обернена додаванню:$(D^1 - D^2) + D^2$=D^1.

Нульовий вектор має властивість: $D + \vec{0}$= D.

З усього вище сказаного, можна зробити висновок, що представлення ІР у вигляді (1) є вектором, а множина ІР складає M-вимірний векторний простір.

Впровадження динаміки в *М*-вимірний пошуковий векторний простір

ІР мають атрибутивними, прагматичними і динамічними властивостями. Атрибутивні – це ті властивості, без яких інформація не існує. Прагматичні властивості характеризують ступінь корисності інформації для користувача, споживача і практики. Динамічні властивості характеризують зміну характеристик ІР в часі.

Найважливішими серед атрибутивних властивостей ІР є дискретність і неперервність. Дискретність виявляється в тому, що в ІР вміщені відомості, знання – дискретні, тобто характеризують окремі фактичні дані, закономірності та властивості досліджуваних об'єктів, які поширюються у вигляді різних повідомлень. ІР, як повідомлення, в яких відображена інформація, мають властивість зливатися з уже зафіксованими і накопиченими раніше, тим самим сприяючи поступальному розвитку і накопиченню. У цьому знаходить своє підтвердження неперервність ІР.

Прагматичні властивості ІР виявляються в процесі використання інформації, відображеної в них. У першу чергу, до даної категорії властивостей відносять наявність змісту і новизни інформації, що характеризує переміщення інформації в соціальних комунікаціях і виділяє ту її частину, яка нова для споживача. Корисною називається інформація, що зменшує невизначеність відомостей про об'єкт. Властивість кумулятивності характеризує накопичення і зберігання ІР.

Динамічні властивості ІР характеризують розвиток ІР в часі. З’являються нові ІР, інші втрачають актуальність – це кількісно відображається на самій моделі ІР.

Втрата з часом інформаційними ресурсами своєї цінності і корисності називається старінням.

Врахування старіння інформації має велике значення при аналітичних дослідженнях, створенні інформаційних продуктів типу інформаційних портретів, основних сюжетів подій, ранжируванні результатів роботи інформаційно-пошукових систем. Навіть наближена оцінка швидкості старіння ІР має величезну практичну цінність, оскільки спонукає надавати більшої значущості актуальним ІР [3].

Старіння ІР проявляється в тому, що постійно виникають нові ІР, нові джерела, які містять більш повну, точну, достовірну інформацію.

При цьому складність використання закономірностей старіння ІР складається з різниці зменшення їх використання в різних предметних областях і для різних тимчасових періодів. Ступінь старіння інформації неоднакова для ІР різних видів і тематик. На швидкість старіння різною мірою впливає дуже багато факторів. Особливості старіння ІР пов'язані з тенденціями розвитку кожного тематичного напрямку. Для того, щоб кількісно оцінити швидкість старіння ІР, Р. Бартон і Р.Кеблер по аналогії з періодом напіврозпаду радіоактивних речовин також ввели поняття «напівперіода життя» наукових статей. Напівперіод життя в їх розумінні – це час, впродовж якого була опублікована половина всіх використовуваних в даний час документів щодо обраної події або явища. Бартон і Кеблер визначили періоди напіврозпаду публікацій з фізики – 4,6 року, з математики – 10,5, геології - 11,8.

Часто використовується модель Мальтуса. Перевагою даної моделі є те, що рівняння Мальтуса має точне рішення у вигляді простої і зручної функції – експоненти, але з точки зору інтерпретації результатів вона виглядає досить сумнівною. Головною проблемою слід вважати, що експонента є монотонно зростаючою функцією, отже, принципово не може описувати процеси, які за своєю природою повинні мати локальні екстремуми, але для великої кількості ІР модель Мальтуса є коректною[3].

Розглянемо модель ІР (1), де для *k*-оготерміну *i*-ого ІР вага w_{ik} визначається формулою (4). Дана формула є добутком стаціонарної складової *TF* і динамічної *IDF*. Тоді, спираючись на модель Мальтуса, отримаємо [4]:

$$w_{ik} = TF_{ik} * IDF_k * e^{-\alpha_c(T_i - T_{i0})} \quad (5)$$

Де i – номер ІР в інформаційному потоці або сховищі;
k – номер терміну в словнику;
t_{ik} – вага k-ого терміна в і-ому ІР;
TF_{ik} – локальна частота k-ого терміну в і-ому ІР визначається формулою (2);
IDF_k – інверсія частоти, з якою деякий термін зустрічається в інформаційному потоці, визначається формулою (3);
A_C – коефіцієнт напіврозпаду актуальності ІР, віднесеного до класу C, визначається експертним шляхом, для кожного класу окремо;
C – клас ІР;
T_i – тривалість часу існування i-ого ІР;
T_{i0} – час виникнення i-ого ІР.

Припустимо, що на відрізку часу $[t_0,t_k]$, згідно з деякими закономірностями, в сховищі з’являється до ІР. На осі часу моменти публікації окремих ІР позначимо як $t_1, t_2, \ldots, t_k$ ($t_0 \le t_1 \le t_2 \le \le \ldots \le t_k$). Інформаційним потоком будемо називати процес $N(t)$, реалізація якого характеризується кількістю ІР, опублікованих в інтервалі (t_0, t). Згідно з експоненціальною моделлю інформаційних потоків:

$$N(t) = N_0 e^{\lambda(t - t_0)}, \quad (6)$$

Де $N(t)$ – кількість ІР в інформаційному потоці в прогнозованому часі;
N_0 – кількість ІР в інформаційному потоці початковий час;
t –час;
t_0 – початковий час;
λ – середня відносна зміна інтенсивності інформаційного потоку: $\lambda(t_i) = \frac{N(t_i) - N(t_{i-1})}{N(t_{i-1})}$.

Відповідно до формули (6) динаміка ІР в інформаційному потоці опишеться:

$$w_{ik} = \frac{m_{ik}}{M_i} ln\left(\frac{N_o e^{\lambda(t - t_0)}}{N_{ok} e^{\lambda_k(t - t_0)}}\right) = \frac{m_{ik}}{M_i}\left[N_0 \ln\left(e^{\lambda(t - t_0)}\right) - N_{0k} \ln\left(e^{\lambda_k(t - t_0)}\right)\right] =$$

$$= \frac{m_{ik}}{M_i}\left[N_0 \lambda(t - t_0) - N_{0k} \lambda_k (t - t_0)\right].$$

Взагалі, вага w_{ik} *k*-ого терміну *i*-ого ІР буде сумою формул (5) і (6).

Висновок

Вперше алгебраїчний підхід до текстових інформаційних ресурсів застосував Дж. Солтон. При цьому багато фахівців даної галузі науки обережно відносяться до такого підходу, посилаючись на те, що немає вагомого обґрунтування подання тексту як вектора. Але представлення (1) задовольняє всім векторним аксіомам, що доводить: текстовий ІР може подаватися у векторному вигляді. Це дає формальне право на застосування алгебраїчного і геометричного апарату для побудови методів та алгоритмів класифікації, розпізнавання й пошуку текстової інформації.

Текстовий ІР є динамічним об’єктом, тому що актуальність інформації, поданої в цих ресурсах змінюється в часі, як і змінюється весь портрет електронного сховища. Відповідно координати вектора, який представляє ІР, є функціями часу. Таке представлення доцільне, оскільки воно позбавляє необхідності кожного разу перераховувати координати ІР, що впливає на ефективність роботи з електронними сховищами, які, на сьогодні, в своїй базі можуть нараховувати величезну кількість ІР.

Література

1. Губин М.В., Морозов А.Б. Влияние морфологического анализа на качество информационного поиска. (http://rcdl.ru/doc/2006/paper_67_v2.pdf).
2. Вулих Б.З. Введение в функциональный анализ. – М.: «Наука», 1967. – 416с.
3. Ландэ Д.В., Фурашев В.Н., Брайчевский С.М., Григорьев А.Н.Основы моделирования и оценки электронных информационных потоков. - К.: ООО "Инжиниринг ", 2006. – 90 с.
4. Ландэ Д.В. Основы интеграции информационных потоков. Монография. – К.: ООО "Инжиниринг ", 2006. – 240 с.

Literatura

1. Gubin M.V., Morozov A.B. Vliyaniemorfologicheskogoanalizanakachestvoinformatsionnogopoiska. (http://rcdl.ru/doc/2006/paper_67_v2.pdf).
2. Vulih B.Z. Vvedenie v funktsionalnyiyanaliz. – M.: «Nauka», 1967. – 416s.
3. Lande D.V., Furashev V.N., Braychevskiy S.M., Grigorev A.N. Osnovyimodelirovaniya i otsenkielektronnyihinformatsionnyihpotokov. - K.: OOO "Inzhiniring ", 2006. – 90 s.
4. Lande D.V. Osnovy iintegratsii informatsionnyih potokov. Monografiya. – K.: OOO "Inzhiniring", 2006. – 240 s.

RESUME

Duchnovska K.
Formation of the research dynamic vector space

Recently there has been the accumulation of arrays of specialized and unstructured text in formation resources in the Internet. Access to them provided information retrieval systems (IRS). IRS algorithms based on algorithmsof vector algebra. These algorithms included a support vector machine, nearest neighbor, naive Bayesian classifier, latent semantic indexing, etc. At same time, many occurred doubts that the text can be represented by a vector. Justification for this representation gave the right to the use of these and other algebraic algorithms. Proof ofthe vector representationof the text is based on seven axiomsof a vector space. It is commutative, associative vector addition and distributive with respect to the amount, associativity of the product, the product to 0 and 1. Proof that all vector axioms are satisfied, it follows from the physical properties ofthe text.

Today textis considered as a static constant in the algorithms of retrieval system. But the information that is supplied the text is dynamic. Changing the number of information resources on various topics leading to a change in the frequency characteristics of the text. Permanent conversion of these characteristics will not bean effective solution to this problem. This task is solved by representation text as vector, whose coordinates depend on time. Such dependenceis derived from Malthus population model. Because of this implementation, the IRS will work with more relevant characteristics of text information resources.

Духновська К.К.

Формування пошукового динамічного векторного простору

Останнім часом, спостерігається накопичення масивів спеціалізованих і неформалізованих текстових інформаційних ресурсів у глобальній мережі *Internet*. Доступ до них забезпечують інформаційно-пошукові системи (ІПС). Алгоритми роботи ІПС базуються на алгоритмах векторної алгебри. До таких алгоритмів належать: метод опорних векторів, метод найближчого сусіда, наївний байєсовський класифікатор, латентно-семантичне індексування і т.д. При цьому у багатьох виникають великі сумніви, що текст може представлятися вектором. Обґрунтування цього подання дає право на застосування даних та інших алгебраїчних алгоритмів. Доведення векторного представлення тексту базується на семи аксіомах векторного простору. Це є комутативність і асоціативність додавання векторів і дистрибутивність відносно суми, асоціативність добутку, добуток на 0 і на 1. Доведення того, що всі векторні аксіоми виконуються, випливає з фізичних властивостей тексту.

На сьогодні, в алгоритмах роботи ІПС, текст розглядається як статична стала. Але інформація, яка подається цим текстом, є динамічною. Зміна кількості інформаційних ресурсів з різної тематики призводить до зміни частотних характеристик тексту. Постійний перерахунок цих характеристик не буде ефективним вирішенням цієї задачі. Така задача вирішується шляхом подання тексту у векторному вигляді, координати якого залежні від часу. Ця залежність виводиться на основі моделі народонаселення Мальтуса. Унаслідок такого впровадження, ІПС буде працювати з більш актуальними характеристиками текстових інформаційних ресурсів.

Надійшла до редакції 03.07.2015

